Finite dimensional approximation and Newton-based algorithm for stochastic approximation in Hilbert space

Ankur A. Kulkarni [a,1], Vivek S. Borkar [b,2]

[a] *Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, U.S.A.*

[b] *School of Technology and Computer Science, Tata Institute of Fundamental Research, Mumbai, India.*

Abstract

This paper presents a finite dimensional approach to stochastic approximation in infinite dimensional Hilbert space. The problem motivated by applications in the field of stochastic programming wherein we minimize a convex function defined on a Hilbert space. We define a finite dimensional approximation to the Hilbert space minimizer. A justification is provided for this finite dimensional approximation. Estimates of the dimensionality needed are also provided. The algorithm presented is a two time-scale Newton-based stochastic approximation scheme that lives in this finite dimensional space. Since the finite dimensional problem can be prohibitively large dimensional, we operate our Newton scheme in a projected, randomly chosen smaller dimensional subspace.

Key words: Stochastic approximation, Hilbert spaces, stochastic programming, convex optimization, random projection.

1 Introduction

Let $(\Omega, \mathscr{F}, \mu)$ be a probability space, ξ be a random variable taking values in an open bounded set $B \subset \Re$ and $h : \Re \times B \to \Re$ be a non-negative measurable function satisfying the following assumption.

Assumption 1.1 *B has the cone property*[3] *. For each $z \in B$ the z-section of h, viz., $h(\cdot, z)$ is strictly convex and twice continuously differentiable and has a (perforce unique) minimum.*

Denote by $\mathscr{H}$, the Hilbert space of $B \to \Re$ functions induced by the inner product

$$\langle f, g\rangle = \int_\Omega f(\xi(\omega))g(\xi(\omega))d\mu.$$

We are interested in an algorithm for minimizing h in the Hilbert space in a certain sense. The Hilbert space minimizer of h is defined as follows.

Definition 1.1 ($\mathscr{H}-$optimal minimizer) *A random variable $f^* \in \mathscr{H}$ is termed an $\mathscr{H}-$optimal minimizer of h if*

$$\int_\Omega h(f^*(\xi(\omega)), \xi(\omega))d\mu \;\; \leq \int_\Omega h(f(\xi(\omega)), \xi(\omega))d\mu$$

for all f in $\mathscr{H}$.

Equivalently,

$$\inf_{f\in\mathscr{H}} \mathbb{E}[h(f)] := \inf_{f\in\mathscr{H}} \mathbb{E}[h(f(\xi(\omega)), \xi(\omega))] = \mathbb{E}[h(f^*)].$$

Let $\mathscr{H}$ be spanned by a complete orthonormal basis $\Phi = \{\varphi_i\}_{i\in\mathbf{N}}$. (This is always possible if $\mathscr{H}$ is separable, i.e., has a countable dense set.) Definition 1.1 can be rewritten by taking $f^*(z) = \sum_{i=1}^{\infty} x_i^*\varphi_i(z)$ for each $z \in B$. We say that $x^* = \{x^*(i)\}_{i\in\mathbf{N}}$ is an $\mathscr{H}-$optimal minimizer of h if

$$\mathbb{E}\left[h\left(\sum x^*(i)\varphi_i\right)\right] \leq \mathbb{E}\left[h\left(\sum x(i)\varphi_i\right)\right] \qquad (1)$$

Email addresses: akulkar3@uiuc.edu (Ankur A. Kulkarni), borkar@tifr.res.in (Vivek S. Borkar).

[1] Work done while visiting School of Technology and Computer Science, Tata Institute of Fundamental Research, Mumbai, India.

[2] Work supported in part by a grant from General Motors India Lab.

[3] B has the cone property if there exists a finite cone $\mathscr{C}$ such that each point $x \in B$ is the vertex of a finite cone $\mathscr{C}_x$ contained in B and congruent to $\mathscr{C}$. See page 66 [1].

for all $x(i) \in \Re, i \in \mathbf{N}$, such that $\sum_{i\in\mathbf{N}} x(i)^2 < \infty$. Throughout this paper we shall use f^* and $\{x^*(i)\}$ interchangeably.

Suppose the values of $h(\cdot,\cdot)$ are observable only in a noise-corrupted form and that the value of the second argument is generated through samples that we cannot control. The task we accomplish in this paper is (a) defining an approximation to f^* and (b) develop an algorithm whose iterates asymptotically approximate f^*. The algorithm is a two time–scale stochastic approximation iteration based on a novel use of the Newton method and subspace minimization.

The important contributions of this article are as follows. Algorithmically, finding f^* either in the sense of Definition 1.1 or (1) is not possible in any computer–coded scheme. Thus defining a *finite dimensional approximation* to f^* that is soundly justifiable is the first significant contribution of this paper. Importantly, with our analysis we can provide a quantitative measure of the goodness of the approximation. The major contribution of this paper is a novel, implementable approximation algorithm for infinite dimensional stochastic approximation. The nature of this problem, minimization in Hilbert space, is quite different from that commonly tackled in stochastic approximation literature. It is motivated from the field of *stochastic programming* (surveyed below) where such problems are natural and have been widely studied. To our knowledge there exists no work that addresses stochastic programming via stochastic approximation. Yet another contribution of this paper lies in bridging this gap.

The paper is organized in the following fashion. Section 2 covers the background of both stochastic programming and stochastic approximation. Section 3 is devoted to justifying a finite dimensional approach. Section 4 presents the algorithm, 5 discusses convergence and the paper concludes in section 6.

2 Background

2.1 Stochastic programming

Let ξ be a random variable on a probability space $(\Omega, \mathscr{F}, \mu)$. The general stochastic nonlinear program is as stated below.

$$\begin{array}{lll} \text{SNLP} & \min\limits_{x,y} & f(x) + \mathbb{E}\left[h(x, y(\xi(\omega)), \xi(\omega))\right] \\ & & u(x) = 0 \\ \text{s. t.} & & a(x, y(\xi(\omega)), \xi(\omega)) = 0 \\ & & b(y(\xi(\omega)), \xi(\omega)) = 0 \\ & & x, y(\xi(\omega)) \geq 0, \qquad \forall\, \omega \in \Omega \end{array}$$

This formulation emerged out of Dantzig's model for decision making under uncertainty [11] (also independently suggested by Beale [3]). Usually a simplification is made in the above problem as:

$$\begin{aligned} h(x, y(\xi(\omega)), \xi(\omega)) &\mapsto h(y(\xi(\omega)), \xi(\omega)) \\ a(x, y(\xi(\omega)), \xi(\omega)) &\mapsto a(x, \xi(\omega)) + d(y(\xi(\omega)), \xi(\omega)) \end{aligned}$$

Solving (SNLP) amounts to finding a deterministic variable x and a random variable y. Finding x is routine and can be done using any conventional optimization techniques. Solving for a function y makes this problem challenging and unique. The canonical problem that is usually used to motivate this model is the "news vendor problem" which we present below. The reader may consult [5] for a thorough introduction to stochastic programming.

2.1.1 News vendor problem

On a given day, a news vendor buys x newspapers at a cost $c(x)$ before the demand for newspapers is known. The newspapers are sold after the materialization of demand, $d_{\xi(\omega)}$, which differs according to scenario (or sample point) $\omega \in \Omega$. Sales, also dependent on ω and denoted by $y(\xi(\omega))$, result in a revenue $q_{\xi(\omega)}(y(\xi(\omega)))$. The unsold newspapers, $w(\xi(\omega))$ $(= x - y(\xi(\omega)))$, are returned back to the supplier at a rate $r_{\xi(\omega)}$. The decision x is called the first stage decision and the tuple $(y(\xi(\omega)), w(\xi(\omega)))$ constitutes the second stage decision for scenario ω. If we assume risk neutrality of the newsvendor, then our objective is to maximize the newsvendor's expected profit. The newsvendor looks to optimize his profit over the two stages by finding a suitable x and a collection $(y(\xi(\omega)), w(\xi(\omega)))_{\omega\in\Omega}$ so as maximize the expected profit, subject to the constraints of demand.

$$\begin{array}{lll} \text{NV} & \min\limits_{x,y,w} & c(x) - \mathbb{E}\left[q_{\xi(\omega)}(y(\xi(\omega))) + r_{\xi(\omega)} w(\xi(\omega))\right] \\ & & y(\xi(\omega)) + w(\xi(\omega)) = x \\ \text{s. t.} & & y(\xi(\omega)) \leq d_{\xi(\omega)} \\ & & x, y(\xi(\omega)), w(\xi(\omega)) \geq 0, \quad \forall \omega \in \Omega. \end{array}$$

A popular direction of research in stochastic programming has been via the assumption of finite Ω. For finite Ω, the problem NV is merely a large nonlinear optimization problem in variables

$$(x, y(\xi(1)), w(\xi(1)), \ldots, y(\xi(|\Omega|)), w(\xi(|\Omega|))),$$

but with a nice structure. Most of previous research has been directed towards exploiting this structure to generate algorithms that are scalable with respect to $|\Omega|$. This direction of work suppresses the *stochasticity* of stochastic programming, which is indeed its most interesting aspect. The earliest work in stochastic programming assumed an infinite probability space [34,25] and laid the groundwork by giving meaning to 'optimality' under uncertainty. Since then, this question has only recently been revisited in [18]. Another approach to solving stochastic programs with infinite Ω has been by using sample average approximations (or empirical expectation) to the $\mathbb{E}[\,\cdot\,]$ term in the objective [27,28,30,16,29].

It can be argued that applying stochastic approximation to stochastic programming is a sensible endeavour. Posing the news vendor's problem as above implicitly assumes the knowledge of functional form of $q_{\xi(\omega)}, \forall\, \omega \in \Omega$ on the part of the news vendor. In reality a news vendor *learns* the form of his profit curve through his experience of past scenarios – samples of demand which are exogenously controlled. The inspiration for applying stochastic approximation to decision making under uncertainty arises from this very standpoint. Using stochastic approximation we allow the learning of the objective function by the decision maker and thus solve the stochastic program.

2.2 Stochastic approximation

The typical approximation scheme to find the extremum of a function g follows the iteration

$$x_{n+1} = x_n + a_n[\nabla g(x_n) + M_{n+1}], \tag{2}$$

where M_n is a martingale difference sequence and the steps a_n satisfy

$$\sum a_n = \infty \qquad \sum a_n^2 < \infty.$$

Robbins and Monro first introduced the stochastic approximation procedure on the Hilbert space $\Re$ [24] to locate the deterministic zero of a function using its noisy measurements. An alternative stochastic approximation scheme was presented by Kiefer and Wolfowitz [17] with a finite difference approximation replacing the term $\nabla g(x_n)$. Apart from being able to deal with noisy measurements, stochastic approximation schemes offer several other advantages. Stochastic approximation methods are robust; in the sense that they have very good convergence properties. The first scheme of Robbins and Monro showed mean square convergence. Stronger convergence results were subsequently obtained by Wolfowitz [35] and Blum [6]. Stochastic approximation is also light on memory usage, requiring the storage of only the previous iterate. From the point of view of analysis of the algorithm, under certain conditions the stochastic approximation algorithm is known to asymptotically resemble the behavior of the trajectory of the ODE

$$\dot{x}(t) = \nabla g(x). \tag{3}$$

Hence the iteration in (2) can be analyzed conveniently using the corresponding ODE in (3). The reader is invited to see chapter 1,2 of [7] for a quick summary.

An alternative iteration to (2) is to use a "Newton-type" approach; an approach that we shall adopt in a certain form that will be made clear later. The Newton iteration looks like this:

$$x_{n+1} = x_n - a_n\left[\nabla^2 g(x_n)^{-1}\nabla g(x_n) + M_{n+1}\right].$$

These methods have also received considerable attention in literature. Their drawback/unattractiveness lies in the $\mathcal{O}(N^2)$ computations needed for Hessian calculation. Hence research in Newton methods has largely followed the Kiefer-Wolfowitz regime via attempts to reduce the computational burden. A fairly extensive summary and an idea of Newton-type approaches is available in [4] and the references therein.

Since Robbins-Monro and Kiefer-Wolfowitz, finite dimensional stochastic approximation in Euclidean space has been a topic of copious theoretical and applied research. Infinite dimensional stochastic approximation is not as rich in its history as its finite dimensional cousin. The reader may see Dvoretsky [13] for the earliest work and Révész [22,23], Walk [32,33] for some related work.

There are broadly two approaches to infinite dimensional stochastic approximation: parametric and nonparametric. The nonparametric or abstract approach applies (2) on objects $x_n \in \mathscr{H}$ that have no parametric specification. Such an approach has the obvious deficiency of being inapplicable to any realistic computer implementation, but is immune to misspecification of parameters. The interested reader may look at [9] for a discussion on the pros and cons of parametric and nonparametric approaches. The convergence analysis for nonparametric stochastic approximation follows from analysing the related $\mathscr{H}-$valued ODE as in [8] or by using probabilistic inequalities as in [22,23].

Alternatively, one may follow a parametric approach using a complete basis as in Eq (1). A popular idea in such a pursuit has been to use a 'sieve' type approach. This idea applies the classical Robbins-Monro (or Kiefer-Wolfowitz) technique to nested finite dimensional subspaces of $\mathscr{H}$ of growing dimension. See Goldstein [14], Nixdorf [19] and Chen and White [9] for examples of such a modus operandi. Any algorithm with perpetually growing bases also suffers from the problem of requiring infinite storage space; and is thus practically unimplementable. Of course one may choose to solve the problem approximately by limiting the size of the subspace to be searched in by *a priori* selecting finitely many basis vectors φ_i and then perform the classical stochastic approximation procedure on Eq (1) on finitely many x_i. But usually it is difficult to justify an a priori knowledge of adequate finitely many basis vectors in infinite dimensional problems – indeed one of our contributions is providing such a justification.

Here we propose to resolve the issue of implementability in infinite dimensional stochastic approximation (an issue which is also inherited by *large* finite dimensional stochastic approximation) by obtaining an N dimensional approximation to (1), $N < \infty$, via a stochastic approximation scheme that runs mostly in k dimensions, $k \approx \mathscr{O}(\log N\epsilon^{-2})$, where ϵ is a degree of accuracy. We now provide a precise definition of this concept and a description of the algorithm.

3 The stochastic approximation algorithm

Recall our intention of incorporating learning in the news vendor problem via exogenously controlled samples. Suppose we are provided with a stream of i.i.d. samples $\{\xi_n\}$, where for each n, $\xi_n : \Omega \to B$ is measurable.

The iterates of our algorithm live in a finite dimensional subspace of $\mathscr{H}$. A justification for this relaxation follows in the next section; here we outline the approach. Suppose for the moment that we are provided with a very large finite set of basis vectors $\widehat{\Phi} = \{\varphi_i\}_{i=1}^N$. From here on $N = |\widehat{\Phi}|$. Let $\widehat{\mathscr{H}} = \left\{\sum_{i=1}^N \alpha_i\varphi_i \mid \alpha_i \in \Re\right\}$. The stochastic approximation is to ensue in $\widehat{\mathscr{H}}$. To simplify matters we use the notation

$$\widehat{h} : \Re^N \to \Re \text{ where } \widehat{h}(x) = h\left(\sum_{i\leq N} x(i)\varphi_i(\xi(\omega)), \xi(\omega)\right).$$

Definition 3.1 (Finite dimensional approximation) *The $N-$dimensional approximation to f^*, denoted by x^* is defined as the minimizer of* $\mathbb{E}\left[\widehat{h}\right]$.

Let x_n denote the 'current point' of the iteration. We *randomly* select a $k-$dimensional subspace of $\widehat{\mathscr{H}}$ and generate iterates $\{y_n\}$ that minimize a quadratic model of $\widehat{h}$ at x_n *in this subspace.* Simultaneously, but through a small increment, we move x_n to x_{n+1}. Occasionally we stop the iteration in current subspace and proceed with another minimization along a freshly selected random subspace. These fresh selections are made increasingly infrequent as the iteration matures. At any time if the iterates escape a prescribed closed convex bounded set, we project them back. This method can be identified as a variant of the classical Newton-type approach. It differs from the classical in two features:

(1) The space of operation changes from time to time, while it does not in usual Newton-type methods.
(2) Classical methods *wait* for the inner iteration to complete before changing to a new point. We instead run both iterations in tandem but with different stepsizes. A similar idea in the Newton context is present in [4].

The reason for changing the subspace in (1.) is to reduce the computational burden – N can be extremely large and stochastic approximation on all N values can be computationally expensive. Our iteration requires an update and computation of only k values at each step. The reason for choosing a *random* subspace is the lack of any other clues to guide our choice. Ideas for this arise from the field of random projection [31] and we shall heavily employ results obtained from there.

The ideas for (2) are chiefly from chapter 6 of [7]. Since stochastic approximation iterations converge only asymptotically, it is impractical to let the quadratic minimization 'finish' and then shift the current point. The same effect can be simulated via a simultaneous iteration with different stepsizes.

The following proposition follows from Assumption 1.1.

Proposition 3.1 $\mathbb{E}[\widehat{h}(x)]$ *is strictly convex in* x.

Due to the Newton-type approach, the algorithm makes a descent at each step. Suppose $f_0 \in \widehat{\mathscr{H}}$ is the initial point of the iteration. Hence f^* lies in the level set

$$S = \{f \in \mathscr{H} \mid \mathbb{E}[h(f(\xi(\omega)), \xi(\omega))] \\ \leq \mathbb{E}[h(f_0(\xi(\omega), \xi(\omega)))]\}.$$

By Assumption 1.1, $h(\cdot, \xi(\omega))$ has bounded level sets. As a consequence S is closed and bounded. The desired 'solution' $\sum_{i\leq N} x^*(i)\varphi_i$ lies in

$$\widehat{S} = S \cap \widehat{\mathscr{H}}.$$

Let $x \in \Re^N$. Consider

$$\widehat{F}(x) = x - a\left[\nabla\widehat{h}(x)\right], \tag{4}$$

with '∇' denoting the gradient operation with respect to $x = [x(1), \ldots, x(N)]^T$. x^* is then a fixed point of $\mathbb{E}\left[\widehat{F}\right]$. Let x_n be an estimate of x^* and let the linear approximation of $\widehat{F}$ at x_n be

$$F^n(x) = \widehat{F}(x_n) + \nabla\widehat{F}(x_n)(x - x_n). \tag{5}$$

Suppose x^{n*} is the fixed point of F^n. It's easy to see that $x^{n*} - x_n = -\nabla^2\widehat{h}(x_n)^{-1}\nabla\widehat{h}(x_n)$. That is x^{n*} can be likened to a 'Newton step' on $\widehat{h}(x)$ [20]. Our problem in fact allows us more structure. Specifically, observe that,

$$\nabla^2\widehat{h}(x_n) = h''(f)\begin{bmatrix} \varphi_1\varphi_1 & \cdots & \varphi_1\varphi_N \\ & \ddots & \\ \varphi_N\varphi_1 & \cdots & \varphi_N\varphi_N \end{bmatrix} = h''(f)\Psi.$$

Notably, N can be so large that it is still impractical to implement an algorithm in N variables. We apply the ideas of [2] to $\widehat{F}$, whereby, the matrix $\nabla\widehat{F}(x_n)$ is projected on to a dimension $k = \mathcal{O}(\log N\epsilon^{-2})$, ϵ denoting a degree of accuracy. The fixed point of F^n in this reduced space is denoted by y^{n*}. It is shown in [2] that y^{n*} lies close to x^{n*} with a high probability. In the following section we recapitulate these ideas briefly.

3.1 Random projection

This section recalls some background material on random projections, based on Chapter 8 in [31] and [2]. Let M be a $N \times N$ real matrix. We may decompose M using its singular values as

$$M = \sum_1^r \sigma_i u_i v_i^T,$$

where r is the rank of M and u_i and v_i are orthonormal with respect to the usual inner product in $\Re^N$. σ_i are the so called 'singular values' of M. Say $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$. Here we demonstrate a rank k approximation to M, called $\widetilde{M_k}$. Now suppose R is a uniform random $\ell \times N$ matrix ($\ell \geq k$). Denote $P = \sqrt{\frac{d}{\ell}}RM^T$. Using its singular values P can be expressed as

$$P = \sum_1^t \lambda_i a_i b_i^T,$$

where t is the rank of P; $a_i \in \Re^\ell, b_i \in \Re^N$. Let

$$\Pi = \sum_1^k b_i b_i^T \tag{6}$$

Observe that Π is a $N \times N$ matrix with rank k. Let $M_k \in \Re^{N\times N}$ denote the following approximation to M:

$$M_k = \sum_1^k \sigma_i u_i v_i^T.$$

This is the best approximation to M w.r.t. the Frobenius norm: $\|A\|_F := (\sum_{i,j} a_{ij}^2)^{\frac{1}{2}}$ for $A = [[a_{ij}]]$. A *lower rank approximation* (of rank k) to M is taken to be

$$\widetilde{M_k} = \Pi M \qquad \in \Re^{d\times d}$$

The following is Theorem 8.5 in [31].

Theorem 1 ([31]) *Let ϵ be prescribed. If $\ell > C\frac{\log d}{\epsilon^2}$ for large enough C, then with high probability*

$$\|M - \widetilde{M_k}\|_F^2 \leq \|M - M_k\|_F^2 + 2\epsilon\|A_k\|_F^2,$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

The first term on the right is by definition the least possible error w.r.t. this norm. Now suppose we intend to find the zero of $G : \Re^N \to \Re^N$, which is known to be strictly monotone[4]. Let z^* be the zero of G and L be a bound on z^*. Equivalently we may find the fixed point of the map $\widehat{G} = z - G$. In the neighbourhood of z^* consider the linearization of $\widehat{G}$

$$F(z) = z^* + (I - a\nabla G(z^*))(z - z^*).$$

Let $\lambda_{\max}$ be the highest eigenvalue of $\nabla G(z^*)$. $F(\cdot)$ is a contraction if the eigenvalues of $(I - a\nabla G(z^*))$ lie inside the unit circle. It follows that for

$$a < \frac{2}{\lambda_{\max}},$$

F is a contraction (with a contraction factor α, say), and z^* is a fixed point of F. Let Π be a projection operation on $\Re^N$ of rank k for k satisfying the hypothesis of Theorem 1, obtained as above. Take $M = I - a\nabla G(z^*)$; define M_k as above and let

$$\eta = \frac{1}{1-\alpha}\left(\left(\|M_k - M\|_F^2 + 2\epsilon\|M_k\|_F^2\right)^{1/2} L + \|b - \tilde{b}\|\right),$$

where $b = a\nabla G(z^*)z^*$ and $\tilde{b} = \Pi b$. Let $\widetilde{G} = G|_{Range(\Pi)}$. We then have the following result from [2].

Theorem 2 *The zero of $\widetilde{G}$ lies in the η neighbourhood of z^* with a high probability.*

The 'high probability' can be made as close to 1 as possible by a standard boosting procedure. We shall set it to $> 1 - \delta$ for a prescribed $\delta << 1$.

3.2 The algorithm

We now motivate and describe the proposed stochastic approximation scheme. Consider the following stochastic approximation algorithm. Suppose $\{\xi_n\}$ is a sequence of i.i.d. samples of ξ. Let $\{a_n\}$ and $\{b_n\}$ be stepsize sequences satisfying

$$\frac{a_n}{b_n} \to 0,$$

[4] $G : \Re^N \to \Re^N$ is said to be strictly monotone if $(G(x) - G(y))^T(x - y) > 0$ for all $x, y \in \Re^N$

in addition to

$$\sum a_n = \infty, \sum b_n = \infty \text{ and } \sum a_n^2 + b_n^2 < \infty.$$

Define the function $\widehat{F}(x, \xi_n) = x - \rho\left[\nabla h(x, \xi_n)\right]$. Let Π^n be a random projection generated at time n using the theory of the above section. At any iterate x_n, denote

$$F^n(x, \xi_n) = \widehat{F}(x_n, \xi_n) + \nabla\widehat{F}(x_n, \xi_n)(x - x_n),$$

where $\nabla\widehat{F}(x, \xi_n) = I - \rho\left[\nabla^2 h(x, \xi_n)\right]$. The fixed point of $F^n(x, \xi_n)$ is the minimizer of

$$\begin{aligned}\psi_n(x) = h(x_n, \xi_n) &+ \nabla h(x_n, \xi_n)^T(x - x_n) \\ &+ \tfrac{1}{2}(x - x_n)^T\nabla^2 h(x_n, \xi_n)(x - x_n),\end{aligned}$$

the quadratic approximation of h near x_n. We minimize $\psi_n(x)$ in the space of the range of the randomly chosen projection Π_n. Let

$$\widetilde{\psi}_n(x) = \psi_n(x_n + \Pi_n x) \text{ and } \nabla\widetilde{\psi}_n = \Pi_n\nabla\psi_n(x_n + \Pi_n x).$$

Thus this minimization is equivalent to finding the fixed point of $\widetilde{F}^n(x) := \Pi_n\left[F^n(x_n + x, \xi_n) - x_n\right]$.

Algorithm 1. Stochastic approximation scheme

```
1. Set $n = 0$. Choose $\epsilon > 0$, $0 < a < 1$
   Determine $N \in \mathbf{N}$ as guaranteed by Theorem 3.
   Determine $k \geq C\frac{\log N}{\epsilon^2}$
   Pick a random projection $\Pi_0$
2. Select $x_0, y_0 \in \Re^N$ such that $x_0, y_0 \in \text{Range}(\Pi_0)$
3. Select a projection $\Gamma$, that projects iterates on a closed convex bounded set
While not terminated do
    i.   Define function $F^n$
         Generate $k \times N$ random matrix $R$. Generate random projection $\Pi^n$ using Eq (6)
    ii.  Denote the projected function
         $\widetilde{F}^n = \Pi_n\left[F^n(x_n + x) - x_n\right]$
    iii. $x_{n+1} = x_n + a_n y_n$
    iv.  $\Pi_{n+1} = \Pi_n + c_n\left[\Pi^n - \Pi_n\right]$
    v.   $y_{n+1} = y_n + b_n\left[\widetilde{F}^n(y_n, \xi_n) - y_n\right]$
    vi.  $x_n = \Gamma x_n, \quad y_n = \Gamma y_n$
    vii. $n = n + 1$
end
```

Theorem 3 mentioned above is presented in the next section. For each n, $c_n \in \{0, 1\}$ controls the change of the k dimensional subspace. Let $\{c_{n(\beta)}\}$ be the maximal subsequence of *all* 1's in $\{c_n\}$. For each $n \in \{n(\beta)\}$, the algorithm switches to a new randomly selected subspace. For the rest, step **iv** can be replaced by $\Pi_{n+1} = \Pi_n$. Furthermore, $\sum c_n = \infty$, implying that such switches are made ad infinitum. We shall also impose $n(\beta + 1) - n(\beta) \uparrow \infty$ as $\beta \to \infty$, meaning that switches become less frequent as the iteration matures. The latter condition will be made more precise later on. We digress now to provide a theoretical justification for the finite dimensional approximation.

4 Justification for a finite dimensional approach

Recall x^* from Definition 3.1, let $\widehat{f} = \sum_{i\in\widehat{\Phi}} x^*(i)\varphi_i$ and consider the following optimization problems.

$$\text{P} \qquad \begin{aligned} &\min_f \ \mathbb{E}\left[h(f)\right] \\ &\text{s. t.} \quad f \in S \end{aligned}$$

$$\widehat{P} \qquad \begin{aligned} &\min_f \ \mathbb{E}\left[h(f)\right] \\ &\text{s. t.} \quad f \in \widehat{S} \end{aligned}$$

f^* solves (P) and $\widehat{f}$ solves $(\widehat{P})$. Algorithm 1 outputs x^*, and effectively solves $(\widehat{P})$. The question we try to answer in this section is:

"Why is problem $(\widehat{P})$ a suitable approximation to (P)?"

Since the objective functions of both problems are the same, it is enough to assess the constraints of the problems. $\widehat{S} \subset S$; so the following is true.

$$\mathbb{E}\left[h(f^*)\right] \leq \mathbb{E}\left[h(\widehat{f})\right] \leq \mathbb{E}\left[h(f^*|_{i\leq N})\right],$$

where $f^*|_{i\leq N}$, resp. $f^*|_{i>N}$ is the projection of f^* to $Range\{\varphi_i, i \leq N\}$, resp. its orthogonal complement.

Let $\mathbb{P}$ be any probability measure on $\mathscr{H}$ that defines a prior belief on where f^* lies in $\mathscr{H}$. Specifically we may choose $\mathbb{P}(E) = 0$ if $E \cap S = \emptyset$. $\mathbb{P}$ is said to be *tight* if for all $\delta > 0$ there exists $E_\delta \subset \mathscr{H}$, E_δ compact, such that $\mathbb{P}(\mathscr{H}\backslash E_\delta) < \delta$.

The following is well known property of metric spaces.

Lemma 4.1 ([21], Theorem 3.2, page 29) *If X is a complete separable metric space, every probability measure on X is tight.*

Since $\mathscr{H}$ is a complete separable metric space with respect to $\|\cdot\|$,

$$\mathbb{P}(f^* \in E_\delta, E_\delta \text{ compact }) > 1 - \delta.$$

The next theorem provides a characterization of compact sets in $\mathscr{H}$.

Theorem 3 *Let $\mathcal{A} \subset \mathscr{H}$ be bounded. $\mathcal{A}$ is relatively compact iff*

$$\forall\, \epsilon > 0 \quad \exists\, N_\epsilon \in \mathbf{N} \quad s.t. \quad \sum_{i>N_\epsilon} \langle f, \varphi_i \rangle^2 < \epsilon \;\; \forall\, f \in \mathcal{A}.$$

Proof : Recall that $\mathcal{S} \subset \mathscr{H}$ is relatively compact iff every sequence in $\mathcal{S}$ has a convergent subsequence.

“ $\Longleftarrow$ ”. Let $\{f_n\}_{n\in\mathbf{N}}$ be a sequence in $\mathcal{A}$. Fix $\epsilon > 0$ and $N_\epsilon \in \mathbf{N}$ guaranteed by the condition of the theorem. We thus have

$$\forall\, n \in \mathbf{N} \qquad \sum_{i>N_\epsilon} \langle f_n, \varphi_i \rangle^2 < \epsilon \tag{7}$$

Let $f_n(i) = \langle f_n, \varphi_i \rangle$. Due to boundedness, there exists a subsequence $\{f_{n(k)}\}_{k\in\mathbf{N}}$ such that the sequence of real numbers $\{f_{n(k)}(1)\}_{k\in\mathbf{N}}$ converges. Let the limit be $f(1)^*$. One can then find a subsequence of this subsequence such that

$$\Re^2 \ni \{f_{n(k(j))}(1), f_{n(k(j))}(2)\}_{j\in\mathbf{N}} \rightarrow [f(1)^*, f(2)^*].$$

Extending this ‘diagonal argument’ [5] further one can find $\mathscr{H} \ni \bar{f}^* = \sum f(i)^* \varphi_i$ such that $f_n(i) \rightarrow f(i)^*$ for each i along a common subsequence. By passing to the limit as $n \rightarrow \infty$ along this subsequence, $\bar{f}^*$ also seen to satisfy the condition in (7). We now show that $f_n \rightarrow \bar{f}^*$ in the Hilbert space.

$$\begin{aligned}
\lim_{n\to\infty} \|f_n - \bar{f}^*\|^2 &\leq \lim_{n\to\infty} \|(f_n - \bar{f}^*)|_{i\leq N_\epsilon}\|^2 \\
&\quad + \lim_{n\to\infty} \|(f_n - \bar{f}^*)|_{i>N_\epsilon}\|^2 \\
&\leq \lim_{n\to\infty} \|(f_n - \bar{f}^*)|_{i\leq N_\epsilon}\|^2 + 2\epsilon \\
&\leq 2\epsilon
\end{aligned}$$

On noting that ϵ can be arbitrarily small, the claim follows.

“ $\Longrightarrow$ ”. We prove this by contradiction. Let $\mathcal{A} \subset \mathscr{H}$ be a relatively compact set for which there exists $\epsilon > 0$ such that

$$\forall\, n \in \mathbf{N} \quad \exists f_n \in \mathcal{A} \qquad \sum_{i\geq n} \langle f_n, \varphi_i \rangle^2 \geq \epsilon.$$

By relative compactness, there exists a subsequence $n(j)$ and $f \in \mathcal{A}$ such that $f_{n(j)} \rightarrow f$, i.e. , $\exists\, J \in \mathbf{N}$ such that $\forall \bar{j} \geq J$, $\|f - f_{n(\bar{j})}\|^2 < \epsilon/2$. Let $j \geq J$ be arbitrary.

$$\|f|_{i\geq n(j)}\|^2 \geq \|f_{n(j)}|_{i\geq n(j)}\|^2 - \|(f - f_{n(j)})|_{i\geq n(j)}\|^2 \geq \epsilon/2$$

This holds for all $j \geq J$. Thus for all $j \geq J$, $\sum_{i\geq n(j)} \langle f, \varphi_i \rangle^2 \geq \epsilon/2$. This contradicts

$$\|f\|^2 = \sum \langle f, \varphi_i \rangle^2 < \infty.$$

This completes the proof. □

Fix $\delta > 0$. With high probability f^* solves problem (P_δ)

$$\boxed{\begin{array}{lll} P_\delta & \min\limits_{f} \; \mathbb{E}\left[h(f)\right] & \\ & \text{s. t.} \quad f \in E_\delta, & E_\delta \text{ compact.} \end{array}}$$

Let $\epsilon > 0$, to be chosen later. Using the above theorem we conclude that there exists $N_\epsilon \in \mathbf{N}$ such that $\|f - f|_{i\leq N_\epsilon}| < \epsilon$ for all $f \in E_\delta$. By taking $N = N_\epsilon$, we get with high probability

$$\begin{aligned}
\mathbb{E}\left[h(f^*)\right] \leq \mathbb{E}\left[h(\widehat{f})\right] &\leq \mathbb{E}\left[h(f^*|_{i\leq N})\right] \\
\text{and} \quad \|f^* - f^*|_{i\leq N}\| &< \epsilon
\end{aligned}$$

By continuity of h, it follows that we can choose ϵ and hence N so that $\mathbb{E}\left[h(\widehat{f})\right]$ is arbitrarily close to $\mathbb{E}\left[h(f^*)\right]$ with high probability.

4.1 An estimate of N

We now provide an estimate of $N = N_\epsilon$ under the further qualification that $f^* \in W^{1,2}(\overline{B}) \subset \mathscr{L}^2$. Let ϵ, δ be as chosen at the end of the previous section.

Assumption 4.1 *$f^* \in V \subseteq E_\delta \cap W^{1,2}(\overline{B})$ such that V is closed and bounded w.r.t $\|\cdot\|_{1,2}$*

For any metric space X and any $U \subset X$, let $\mathcal{N}(U, \epsilon, \|\cdot\|_X)$ be the covering number: the least number of balls of radius ϵ with respect to $\|\cdot\|_X$ that can cover U. Suppose we can cover V with balls of radius ϵ w.r.t. $\|\cdot\|$ and let $f_i^o, i \leq \mathcal{N}(V, \epsilon, \|\cdot\|)$ be the centers of these balls. Then for each $f \in V$, there exists $g \in span(\{f_i^o : i \leq \mathcal{N}(V, \epsilon, \|\cdot\|)\})$ such that $\|f - g\| < \epsilon$. Thus we may take $N = N_\epsilon = \mathcal{N}(V, \epsilon, \|\cdot\|)$.

Several definitions and results used below are from [1]. We give a bound on $\mathcal{N}(V, \epsilon, \|\cdot\|)$ using [10]. Since $\|\cdot\|$ is dominated by $\|\cdot\|_\infty$,

$$\mathcal{N}(V, \epsilon, \|\cdot\|) \leq \mathcal{N}(V, \epsilon, \|\cdot\|_\infty).$$

[5] That is, we construct a nested sequence of subsequences each containing the next, such that for each k, the first k components of the k^{th} subsequence converge. Then by picking the k^{th} element of the k^{th} subsequence, we have a subsequence all of whose components converge. See, e.g., [26].

It is known that the embedding

$$\mathscr{E} : W^{1,2}(\overline{B}) \hookrightarrow C_b(\overline{B})$$

is compact (Rellich-Kondrachov Theorem in [1]). Define the m^{th} *entropy number* of a metric space X to be

$$\begin{aligned} e_m(X) = \inf \{ & \varepsilon > 0 \mid \exists \text{ closed balls} \\ & D_1, \ldots, D_{2^m - 1} \text{ with radius } \varepsilon \text{ covering } X \} \end{aligned}$$

Using (4), page 16 of [10], we have that if $\overline{B}$ has a C^∞ boundary, then

$$e_m(\mathscr{E}(\mathscr{B}_1)) \leq \kappa \left(\frac{1}{m} \right)^2,$$

where $\mathscr{B}_R = \{ f \in W^{1,2}(\overline{B}) \mid \|f\|_{1,2} \leq R \}$ and κ is a constant independent of m.

Let $\mathscr{B}_R \supseteq V$ for some R. $\mathscr{B}_R$ is compactly embedded in $C_b(\overline{B})$. Using Proposition 6, page 16, in [10]

$$\begin{aligned} \ln \mathcal{N}(V, \epsilon, \|\cdot\|) &\leq \ln \mathcal{N}(\overline{\mathscr{E}(V)}, \epsilon, \|\cdot\|_\infty) \\ &\leq \ln \mathcal{N}(\overline{\mathscr{E}(\mathscr{B}_R)}, \epsilon, \|\cdot\|_\infty) \\ &\leq \left(\frac{R\kappa}{\epsilon} \right)^{1/2} + 1. \end{aligned}$$

Using the dominance of norms stated above, V is closed with respect to $\|\cdot\|$ and $\ln \mathcal{N}(V, \epsilon, \|\cdot\|) \leq \ln \mathcal{N}(\overline{\mathscr{E}(V)}, \epsilon, \|\cdot\|_\infty)$. If we take the span of the vectors that form the centers of these balls, that will clearly give a finite dimensional approximation with the above error bound. This gives an order of magnitude estimate for the size of the approximating subspace needed.

5 Convergence analysis

5.1 Preliminaries

The algorithm above is a projected simultaneous stochastic approximation iteration in (x_n, y_n) but with different time scales.

$$x_{n+1} = \Gamma\left(x_n + a_n y_n\right) \tag{8}$$
$$y_{n+1} = \Gamma\left(y_n + b_n \left[g(\Pi_n, x_n, y_n) - y_n + M_{n+1} \right]\right), \tag{9}$$
$$\Pi_{n+1} = \Pi_n + c_n \left[\Pi^n - \Pi_n \right] \tag{10}$$

where $g(\Pi_n, x_n, y_n) = \int \widetilde{F}^n(y_n, \xi)\zeta(d\xi)$, where ζ is the law of ξ_n for all n and $M_{n+1} = \widetilde{F}^n(y_n, \xi_{n+1}) - g(\Pi_n, x_n, y_n)$ is a martingale difference sequence by construction. Due to projection, $\{x_n, y_n\}$ stay bounded. Recall that we had set $\{a_n\}, \{b_n\}$ such that $\frac{a_n}{b_n} \to 0$, in addition to the usual conditions on stepsizes of stochastic approximation algorithms. $\{c_n\}$, which we specify later, also satisfies $\frac{c_n}{b_n} \to 0$. The analysis of two time scale algorithms from [7], section 6.1 then allows us to treat $\{x_n\}, \{\Pi_n\}$ as quasi-static and analyze $\{y_n\}$ in isolation treating the former as constant. By the 'o.d.e.' analysis of [7], section 5.4, $\{y_n\}$ has a.s. the same asymptotic behavior as the o.d.e.

$$\dot{y}(t) = g(\Pi, x, y(t)) - y(t) + r(t), \tag{11}$$

where '$r(t)$' is a boundary correction term. If we assume that the vector field $g(\Pi, x, y) - y$ is transversal and pointing inwards at every point of the boundary $\partial\Gamma$ of Γ, this correction term is identically zero. We do so for simplicity, though the condition could be relaxed with some additional technicalities. This reduces (11) to

$$\dot{y}(t) = g(\Pi, x, y(t)) - y(t) \tag{12}$$

5.2 Convergence of fast iteration

We use the following notation.

$$\begin{aligned} \Re^N \ni x^* &: \text{The desired solution} \\ \Re^N \ni x^{n*} &: \text{Fixed point of local approximation} \\ \Re^N \ni y^{n*} &: \text{Fixed point of projected approximation} \end{aligned}$$

where

$$\begin{aligned} \mathbb{E}\left[\nabla h(x^*, \xi_1(\omega)) \right] &= 0, \\ \mathbb{E}\left[F^n(x_n + x^{n*}, \xi_1(\omega)) - x_n \right] &= x^{n*} \\ \mathbb{E}\left[\widetilde{F}^n(y^{n*}, \xi_1(\omega)) | \Pi_n \right] &= y^{n*} \end{aligned}$$

Assumption 5.1 *There exists $K > 0$, such that $\mathbb{E}\left[\|M_{m+1}\|^2 | \mathscr{F}_m \right] < K$ a.s. for each m.*

The following lemma is easy to see.

Lemma 5.1 *$\lambda_{\max}(\xi(\omega))$, the maximum eigenvalue of $\Psi(\xi(\omega))$, is bounded a.s.*

Theorem 4

$$\sup_{n \in \mathbf{N}} \operatorname{ess\,sup}_\Omega \lambda_{\max}(\xi_n(\omega)) h''(\sum x_n(i)\varphi_i, \xi_n) < \infty,$$

and for

$$\rho < \inf_{n \in \mathbf{N}} \operatorname{ess\,inf}_\Omega \frac{2}{\lambda_{\max}(\xi_n(\omega)) h''(\sum x_n(i)\varphi_i, \xi_n)},$$

$\widetilde{F}^n(x, s)$ is a uniform (w.r.t. s, n) contraction in x a.s.

The first claim follows from Lemma 5.1 and the boundedness of iterates. The second claim is an easy consequence of this and Theorem 2 in section 3.1. The above theorem ensures that there exists a unique fixed point y^{n^*} of $g(\Pi, x, \cdot)$.

Theorem 5 $y_m \to y^{n^*}$ *a.s.*

The convergence of (12) to y^{n^*} follows by Theorem 2, p. 126, of [7]. The claim then follows by Theorem 2, p. 15, of [7].

Note that y^{n^*} will be a function of Π, x, say $y^{n^*} = y^{n^*}(\Pi, x)$. What the above means is that $y_n - y^{n^*}(\Pi_n, x_n) \to 0$ a.s. (see [7], section 6.1). Our conditions on $\{c_n\}$ stated later also ensure that $\frac{c_n}{a_n} \to 0$, which in view of the above and section 6.1 of [7], ensures that we can now analyze the iterates $\{x_n\}$ treating Π_n as constant $\approx \Pi$ and $y_n \approx y^{n^*}(\Pi, x_n)$. As a solution to a parametrized quadratic minimization problem stated earlier, $y^{n^*}(\Pi, x)$ will be Lipschitz in x.

As an application of Theorem 2, we get the next theorem.

Theorem 6 y^{n^*} *lies within an* η *neighbourhood of* x^{n^*} *with probability* $> 1 - \delta$.

5.3 Convergence of slow iteration

We shall adapt the arguments of [7], section 4.2, and sketch the convergence arguments in outline. The full details would follow closely the corresponding arguments in [7] and would be excessively lengthy. Note that $\{x_n\}$ has a.s. the same asymptotic behavior as the o.d.e. [6]

$$\dot{x}(t) = y^{n^*}(\Pi, x(t)) = x^{n^*}(x(t)) + c(t), \qquad (13)$$

where $\|c(t)\| < \eta$ with probability $> 1 - \delta$. Suppose the latter holds. Compare this with

$$\dot{\tilde{x}}(t) = x^{n^*}(\tilde{x}(t)). \qquad (14)$$

Eq (14) is simply the Newton's algorithm in continuous time applied to $V(x) = E[\hat{h}(x)]$ restricted to the subspace under consideration (say, X). Thus V itself serves as a Liapunov function for (14). Recall that we are operating in a bounded set, say $\tilde{S} \subset X$. Consider $S' := \tilde{S}-$ the ϵ'-neighbourhood of the unique minimizer of V in X. Fix $T > 0$. Then along any trajectory $\tilde{x}(\cdot)$ of (14) initiated in S' and of duration $\geq T$, V decreases by at least a certain $\Delta > 0$. By a standard argument based on the Gronwall inequality, the trajectory $x(\cdot)$ of (13) with the same initial condition and duration as above, remains in a small tube around the trajectory $\tilde{x}(\cdot)$ whose width (say, κ) can be made arbitrarily small by choosing η small enough. Let the initial condition be x_{n_0}, where n_0 is the instant when the projection Π under consideration was introduced (in particular, $c_{n_0} = 1$). Set $n_1 := \min\{m > n_0 : \sum_{i=n_0}^{m} a_i \geq T\}$ and suppose as above that the projection is not changed till n_1. By the foregoing and the arguments of Lemma 1, p. 12, of [7], it follows that with probability $> 1 - 2\delta$, for n_0 sufficiently large, $x_m, n_0 \leq m \leq n_1$, will remain in a tube of width κ around $x(t), t \geq \sum_0^{n_0} a_i$, therefore in a tube of width 2κ around $\tilde{x}(t), t \geq \sum_0^{n_0} a_i$. For κ sufficiently small, this ensures that $V(x_{n_1}) < V(x_{n_0}) - \frac{\Delta}{2}$. Since V is bounded, say by $K' > 0$, we have

$$E[V(x_{n_1})] \leq (1 - 2\delta)(V(x_{n_0}) - \frac{\Delta}{2}) + 2\delta K'.$$

The r.h.s. is $< V(x_{n_0}) - \frac{\Delta}{4}$ (say) if δ is chosen sufficiently small. It is important to note that in the above, both Δ and K' can be chosen to be independent of Π_n since the iterates are bounded. Thus the above conclusions hold regardless of the specific choice of n_0 from among the $\{n(\beta)\}$.

We now specify our choice of $\{c_n\}$. Recall that $\{n(\beta)\} \subset \{n\}$ is the maximal subsequence along which $c_n = 1$. Thus it suffices to specify $\{n(\beta)\}$. Define it recursively by: $n(0) = 0$ and $n(\beta + 1) := \min\{m \geq n(\beta) : \sum_{k=n(\beta)}^{m} a_m \geq T\}$.

Let $\hat{x}$ be the minimizer of $V(\cdot)$ in the space X. Let $\gamma = \max_{\|x - \hat{x}\| \leq \epsilon'} V(x)$. Then

$$E[V(x_{n(\beta+1)})|x_m, m \leq n(\beta)] \leq V(x_{n(\beta)}) - \frac{\Delta}{4}$$
$$\text{when} \;\; V(x_{n(\beta)}) - V(\hat{x}) > \gamma.$$

By standard arguments (see, e.g., [15]) it follows that $\{x_{n(\beta)}\}$ will a.s. hit the set $V_\gamma := \{x : V(x) \leq V(\hat{x}) + \gamma\}$. We can now adapt arguments of [7], p. 42, leading to Lemma 13 to show that $\{x_n\}$ cannot escape $V_{\gamma+\nu}$ thereafter for a prescribed small ν. Since both γ and ν can be made arbitrarily small, this implies that x_n converges to the unique minimizer of V a.s.

Recall that $\hat{x}$ is a function of Π_n. We expect that as n increases, $\hat{x}$ reaches a region from which x^* is accessible via a Newton step in $\Re^N$, i.e. , $\hat{x} + x^{n^*}(\hat{x}) \approx x^*$. Thus $\{x_n\}$ converges to within η radius of x^*, where η is as in Theorem 2, with z^* therein taken as x^*.

6 Conclusions

This paper has presented a finite dimensional approach to stochastic approximation in Hilbert space with an

[6] Once again we are ignoring the boundary effects by assuming appropriate transversality condition at the boundary for the vector field under consideration. We skip the details.

application to the field of stochastic programming. The algorithm presented was a two time-scale Newton scheme. Acknowledging that the finite dimensional problem can be prohibitively large dimensional, we operated our Newton scheme in a projected, $\mathcal{O}(\log N)$ dimensional subspace. Admittedly, finding the projection as indicated in section 3.1 can be computationally cumbersome. But we believe that exploiting the structure of Ψ and using techniques such as those in [12], this difficulty can be considerably mitigated.

References

[1] R. A. Adams. *Sobolev Spaces.* Pure and Applied Mathematics, 65. A Series of Monographs and Textbooks. Academic Press, New York-San Francisco-London, 1975.

[2] K. Barman and V. S. Borkar. A note on linear function approximation using random projections. *Systems & Control Letters*, 57(9):784–786 2008.

[3] E. M. L. Beale. On minimizing a convex function subject to linear inequalities. *Journal of the Royal Statistical Society*, 17B:173–184, 1955.

[4] S. Bhatnagar. Adaptive Newton-based multivariate smoothed functional algorithms for simulation optimization. *ACM Transactions Modeling and Computer Simulation*, 18(1):1–35, 2007.

[5] J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming: Springer Series in Operations Research.* Springer, 1997.

[6] J. R. Blum. Approximation methods which converge with probability one. *Annals of Mathematical Statistics*, 25:382–386, 1954.

[7] V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint.* Hindustan Book Agency, New Delhi, India and Cambridge University Press, Cambridge, UK., 2008.

[8] X. Chen and H. White. Nonparametric adaptive learning with feedback. *Journal of Economic Theory*, 82(1):190–222, September 1998.

[9] X. Chen and H. White. Asymptotic properties of some projection-based Robbins-Monro procedures in a Hilbert space. *Studies in Nonlinear Dynamics & Econometrics*, 6(1):1–53, 2002.

[10] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1):1–49, 2002.

[11] G. B. Dantzig. Linear programming under uncertainty. *Management Science*, 1:197–206, 1955.

[12] P. Drineas, R. Kannan and M. W. Mahoney. Fast Monte Carlo algorithms for matrices I: approximating matrix multiplication, *SIAM J. Comput.*,36(1):132–157, 2006.

[13] A. Dvoretsky. On stochastic approximation. In *Proceedings of the 3^{rd} Berkeley Symp. Math. Stat. Prob. 1*, pages 39–55, 1956.

[14] L. Goldstein. Minimizing noisy functionals in Hilbert space: An extension of the Kiefer-Wolfowitz procedure. *Journal of Theoretical Probability*, 1(2):189–204, 1988.

[15] B. Hajek. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Advances in Applied Probability*, 14(3):502–525, 1982.

[16] T. Homem-De-Mello. Variable-sample methods for stochastic optimization. *ACM Trans. Model. Comput. Simul.*, 13(2):108–133, 2003.

[17] J. Kiefer and J. Wolfowitz. Stochastic estimation of a regression function. *Annals of Mathematical Statistics*, 23:462–466, 1952.

[18] A. A. Kulkarni and U. V. Shanbhag. Recourse-based stochastic nonlinear programming: properties and Benders-SQP algorithms. *To appear in Computational Optimization and Applications.*

[19] R. Nixdorf. An invariance principle for a finite dimensional stochastic approximation method in a Hilbert space. *Journal of Multivariate Analysis*, 15:252–260, 1984.

[20] J. Nocedal and S. J. Wright. *Numerical Optimization.* Springer Series in Operations Research. Springer-Verlag, New York, 1999.

[21] K. R. Parthasarathy. *Probability Measures on Metric Spaces.* AMS, Providence, 2005.

[22] P. Révész. Robbins-Monro procedure in a Hilbert space and its application in the theory of learning processes, I. *Stud. Sci. Math. Hung.*, 8:391–398, 1973.

[23] P. Révész. Robbins-Monro procedure in a Hilbert space, II. *Stud. Sci. Math. Hung.*, 8:469–472, 1973.

[24] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

[25] R. T. Rockafellar and R. J.-B. Wets. Stochastic convex programming: Kuhn-Tucker conditions. *J. Math. Econom.*, 2(3):349–370, 1975.

[26] J. G. Rosenstein. *Linear Orderings.* Academic Pr, October 1982.

[27] A. Shapiro. Monte Carlo sampling methods. In *Handbook in Operations Research and Management Science*, volume 10: 353–426. Elsevier Science, Amsterdam, 2003.

[28] A. Shapiro and T. Homem de Mello. A simulation-based approach to two-stage stochastic programming with recourse. *Mathematical Programming: Series A*, 81(3):301–325, 1998.

[29] A. Shapiro, J. Linderoth and S. Wright. The empirical behavior of sampling methods for stochastic programming. *Optimization Technical Report 02-01, Computer Sciences Department, University of Wisconsin-Madison*, 2002.

[30] A. Shapiro and H. Xu. Stochastic mathematical programs with equilibrium constraints, modeling and sample average approximation. *Optimization-Online*, 2005.

[31] S. Vempala. *The Random Projection Method.*, *DIMACS series*, volume 65. AMS, Providence, 2004.

[32] H. Walk. An invariance principle for the Robbins-Monro process in a Hilbert space. *Z. Wahrsch. verw. Gebiete*, 39:135–150, 1977.

[33] H. Walk. Martingales and the Robbins-Monro procedure in $D[0, 1]$,. *Journal of Multivariate Analysis*, 8:430–452, 1978.

[34] D. W. Walkup and R. J.-B. Wets. Stochastic programs with recourse. *SIAM Journal of Applied Mathematics*, 15(5):1299–1314, sep 1967.

[35] J. Wolfowitz. On the stochastic approximation method of Robbins and Monro,. *Annals of Mathematical Statistics*, 23:457–461, 1952.